International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Comparative and Explainable Machine Learning Framework for Fake News Detection: A Trust Gap and Cross-Dataset Robustness Analysis

Akash Suri, Aryan Pathania, Divyayush Verma, and Rajat Takkar
Department of Computer Science and Engineering Chitkara University, Punjab, India

Abstract- The proliferation of fake news on social media plat- forms poses significant threats to public discourse and democratic processes. While numerous machine learning approaches have been proposed for fake news detection, limited attention has been given to understanding why different models classify news as fake and whether these explanations are consistent across algorithms. This paper presents a comparative and explainable machine learning framework that addresses two critical research questions: (1) Do different ML models agree on which textual fea- tures indicate fake news? (Trust Gap Analysis), and (2) Do fake news patterns learned from one domain generalize to another? (Cross-Dataset Robustness). We evaluate four classical machine learning algorithms—Logistic Regression, Naive Bayes, Support Vector Machine, and Random Forest—using TF-IDF features on two distinct datasets: ISOT (political news, 44,898 articles) and WELFake (general news, 72,134 articles). Using SHAP (SHapley Additive exPlanations) for model interpretability, we compute Jaccard similarity and Spearman rank correlation to quantify agreement between model explanations. Our results reveal that different models exhibit varying levels of agreement on fake news indicators, with implications for model selection in real- world deployment. Furthermore, cross-dataset analysis identifies "universal" fake news features that generalize across domains versus "topic-specific" features that are domain-dependent. This work contributes a novel analytical framework for evaluating the trustworthiness and generalizability of fake news detection systems.

Keywords- Fake News Detection, Explainable AI, SHAP, Machine Learning, Trust Gap Analysis, Cross-Dataset Robust- ness.

I. INTRODUCTION

Social media platforms have changed the way of spread of information. This transformation has also unfortunately enabled the rapid information spread of fabricated content to mislead people who read it. When false information bluffs as real journalism, it can affect public opinion in a negative way, disturb electoral processes, and even endanger lives [1]. Research published in science demonstrated that false news spread approximately six times faster than accurate news and have a 70% higher probability of being shared [1]. The global health crisis of 2020 brought more attention to this issue, as health authorities struggled to fight what they termed an infodemic of medical misinformation [2].

Fake news doesn't just hurt people, it costs money too. Analysts estimated that false information costs businesses and governments billions annually through market disruptions, public health issues, and resources spent on to verify the same. Researchers responded by developing automated systems that use ML(machine learning) to identify false content [3], [4]. These systems have grown increasingly sophisticated, progressing from simple word-counting approaches to com- plex neural architectures. Yet a fundamental problem persists: most research prioritizes prediction accuracy while neglecting to explain the reasoning behind classifications. This creates several practical challenges:

- **Opaque Decision-Making:** When a system flags content as potentially false, users receive no insight into what trig- gered the classification. This opacity erodes confidence in automated moderation.
- **Inconsistent Reasoning:** Two algorithms might both correctly identify false content but for entirely different reasons. Without examining their reasoning, we cannot determine which approach captures genuine misinforma- tion patterns versus superficial correlations.

- **Limited Generalization:** A classifier trained on political misinformation may fail completely when encountering health-related false claims. Understanding which detection signals transfer across topics remains poorly understood.
- **Susceptibility to Manipulation:** Without knowing what features classifiers rely upon, we cannot assess how easily bad actors might craft content that evades detection.

This research introduces an logical way that moves beyond accuracy measurements to examine how classifiers reason about misinformation. We have two analyses:

Trust Gap Examination: We observe that classifiers who achieve good accuracy often rely on different text-signals. We develop metrics to quantify this divergence, which we call as the "Trust Gap." Understanding where models agree and disagree provides important guidance for deployment decisions and ensemble construction.

Cross-Domain Feature Analysis: We investigate which misinformation signals appear repeatedly across different news categories versus which signals depend on the topic. This distinction has direct implications for building systems that generalize beyond their training domain.

Our specific contributions include:

1. **A Framework for Comparing Model Explanations**: We present methods using SHAP-derived feature attributions, set overlap metrics, and rank correlation to measure agreement between classifier explanations. This enables principled comparison of how different algorithms approach the detection task.
2. **Systematic Cross-Domain Analysis:** We catego- rized important features as either universal (appearing largely/consistently/repeatedly across news domains) or topic-bound (specific to particular content areas). We introduce the Universality Ratio to quantify cross-domain potential.
3. **Large-Scale Empirical Investigation:** We evaluate 4 different algorithms across 2 substantial datasets, one focused on political content (44,898 articles) and another covering general news (72,134 articles), providing solid guidance for practitioners.
4. **Open Implementation:** We release our complete experimental pipeline to enable replication and extension of this work.

The following sections proceed as follows: Section II surveys prior research. Section III details our analytical methods. Section IV describes experimental configuration. Section V presents findings and their interpretation. Section VI offers conclusions and identifies promising research directions.

II. RELATED WORK

We examine three research area relevant to our investigation: computational methodology for pointing out fabricated content, methods for making classifier decisions explainable, and hurdles in transferring detection capabilities over different content domains.

A. Computational false information Detection

Automated identification of fake content has progressed through several analytical phases.

1) **Feature-Engineered Approaches:** Pioneering work in this area focused on manually designed indicators. Castillo et al. [5] investigated reliability markers in social media, finding out some textual and behavioral patterns are closely related with information authenticity. Horne and Adali [19] conducted detailed analysis based on language, finding that fake articles Mostly seem to be articles with fascinating titles and simpler body text.

Researchers have find out and recorded various textual characteristics that distinguish legitimate from fabricated reporting:

- **Word-level patterns:** Vocabulary choices, term repetition, and phrase structures
- **Sentence construction**: Grammatical complexity, clause arrangements, and punctuation tendencies
- **Emotional markers:** Sentiment intensity, subjective language, and persuasive elements
- **Presentation conventions:** Formatting choices, citation practices, and attribution patterns

Ahmed and colleagues [6] showed that if we combine together word sequence features with standard classifiers we can make effective detection systems. Pe´rez-Rosas et al. [20] extended this work, validating that writing style captures meaningful authenticity signals.

2) **Representation Learning Methods:** Neural architectures brought the ability to harvest the features automatically from raw text itself . Convolutional designs

collects localized textual patterns [8], while recurrent structures model dependencies spanning sentences [9].

Pre-trained language models represent the current frontier. Kaliyar et al. [10] demonstrated that calibrating BERT results in strong detection performance by making use of the semantic knowledge that is acquired during pre-training still neural methods possess an interpretability challenge: their predictions emerge from large no of parameters interact- ing in complex ways. When such a system marks content as fabricated, tracing the exact textual characteristic responsible for the result proves difficult. This limitation increases our significance on interpretable classical approaches.

3) **Propagation and Network Analysis:** Complementing content analysis, researchers have examined how information spreads. Misinformation often possess particular sharing pat- terns, user engagement signatures, and network formation dy- namics. These contextual signals provide additional detection that we can make use of beyond textual content alone.

B. Making Text Classifiers Interpretable

The need to understand algorithmic decisions has driven considerable research into explanation methods [11].

- **Transparent Model Architectures:** Certain classifier fam- ilies permit direct interpretation. Linear models assign co- efficients to features that quantify their contribution to the result. Tree-based models trace clear decision paths. These architectures trade some predictive power for transparency.
- **Retrospective Explanation Approaches:** When employ- ing complex models, explanation techniques can shed some light their reasoning after training. Ribeiro et al. [12] intro- duced LIME, which constructs local interpretable approxima- tions around individual predictions. Lundberg and Lee [13] developed SHAP, which finds out feature attributions from game-theoretic principles.

SHAP provides guarantees particularly important for com- parative analysis:

- Attribution values aggregate to match model output (ad- ditivity)
- Missing features contribute nothing (null contribution)
- Features with equivalent influence receive equivalent at- tribution (fairness)

Explanation methods have been applied to misinformation detection. Reis et al. [14] made use of LIME to find out the words driving classifier decisions. Shu et al. [15] de- veloped systems which also generate explanations alongside predictions. These efforts typically examine individual models; our work extends the scope of this research to cross-model comparison.

C. Transferring Detection Across Domains

Deploying detection systems beyond the domain they are trained presents steady problems [16]. Classifiers effective on political content often struggle with health misinformation, entertainment rumors, or financial deception.

Several factors contribute to this problem:

- Topic areas employ distinct terminology and conventions
- Deception strategies may vary across content categories
- Training corpora may encode topic-specific artifacts rather than generalizable patterns

Zhang and Ghorbani [25] reviewed these challenges, ob- serving that systematic study of which signals transfer remains limited. Our cross-domain investigation directly addresses this gap.

D. Open Questions

Prior research leaves several issues unresolved:

- Do classifiers achieving comparable accuracy employ similar reasoning? Systematic measurement of expla- nation agreement across algorithms has not been at- tempted.
- Which detection signals persist across content domains? Studies document performance drops but rarely analyze feature-level transferability.
- How do explanation consistency and domain robustness connect? This relationship awaits investigation.

The framework presented next addresses each question through supporting analytical approaches.

III. METHODOLOGY

This section presents our complete framework for relative and Understandable fake news detection. We standardize the problem, then we will explain each component of our strategy in detail.

A. Problem Formulation

Given a news article x represented as text, the task is to assign a binary label y ∈ {0, 1} where y = 0 indicates valid news and y = 1 indicates fake news. Beyond classification, we are also trying to:

- Identify which features fi contribute most to the final result
- Compare feature relevance across different models
- Analyze feature applicability across datasets

Formally, let M = {M1, M2, ..., Mk} be a set of k trained classifiers, and let DA and DB be two datasets from different realms. For each model Mi and dataset Dj, we calculate :

- Classification performance metrics (accuracy, precision, recall, F1)
- Find Feature importance rankings via SHAP values
- Pairwise model agreement (Trust Gap)
- Cross-dataset feature overlap (Universality Ratio)

B. Text Preprocessing

Raw news articles contain irrelevant texts that dont contribute to they result and can interfere with classification process. We apply a systematic Data Cleaning Sequence to normalize the text while maintaining the core information.

Require: Raw text T
Ensure: Cleaned text T ′
1: T ← lowercase(T) {Normalize case}
2: T ← remove urls(T) {Remove hyperlinks}
3: T ← remove html(T) {Strip HTML tags}
4: T ← remove punctuation(T) {Remove special charac- ters}
5: T ← remove numbers(T) {Remove digits}
6: tokens ← tokenize(T) {Split into words}
7: tokens ← remove stopwords(tokens) {Remove com- mon words}
8: tokens ← {t ∈ tokens : len(t) > 2} {Remove short tokens}
9: T ′ ← join(tokens, " ")
10: return T ′

- **The preprocessing steps are designed to:**
- Reduce noise: URLs, HTML tags, and special characters are irrelevant text without semantic value
- Normalize vocabulary: Lowercasing ensures Trump and trump are treated in same manner
- Focus on content words: Stopword removal help's to give importance to word that carry the real message
- Reduce dimensionality: Removing very short tokens helps to get rid of fake data points

C. Feature Extraction

We use Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction, it find's a sweet spot between how much a term is used in the current document compared to entire dataset.

For a term t in document d within Collection D, the TF-IDF weight is computed as:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \quad (1)$$

where Term Frequency is:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

and Inverse Document Frequency is:

$$\text{IDF}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}| + 1}$$

We limit the vocabulary to the top 5,000 features by document frequency to balance meaningful and computational efficiency. This threshold was chosen based on early experiments showing no fruitful returns beyond this point.

Classification Models

We evaluate four classical machine learning algorithms, chosen for their diversity in learning paradigms and their Adaptability to SHAP-based explanation:

Logistic Regression (LR): A linear model that evaluates the probability of fake news using the logistic (sigmoid) function:

$$P(y = 1|x) = \sigma(\beta_0 + \beta^T x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

The model is trained by reducing the binary cross-entropy loss with L2 regularization:

$$\mathcal{L} = -\sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] + \lambda ||\beta||_2^2$$

Logistic Regression is naturally explainable: the coefficient βj indicates the log-odds change for a unit increase in feature xj.

Multinomial Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with the naive conditional independence assumption:

$$P(y|x_1,...,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,...,x_n)}$$

The class with maximum posterior probability is selected

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

Despite its simplifying assumptions, Naive Bayes time and again performs well on text classification tasks and produce probabilistic analysis.

Support Vector Machine (SVM): A Discriminative Learning Algorithim that finds the optimal hyperplane maximizing the margin between classes. For linearly separable data, SVM solves:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \forall i$$

For non-separable data, we use soft-margin SVM with slack variables ξi:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i$$

We use LinearSVC for resource optimization with high-dimensional TF-IDF features.

Random Forest (RF): An ensemble learning method that builds multiple decision trees using bootstrap aggregating (bagging) and random feature selection:

Predictions are made by majority voting:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_T(x)\}$$

Algorithm 2 Random Forest Training

Require: Training data $\mathcal{D}$, number of trees T, features per split m
Ensure: Ensemble of trees $\{h_1, ..., h_T\}$
1: **for** $t = 1$ to T **do**
2: $\mathcal{D}_t \leftarrow$ bootstrap_sample($\mathcal{D}$)
3: $h_t \leftarrow$ train_tree($\mathcal{D}_t, m$)
4: **end for**
5: **return** $\{h_1, ..., h_T\}$

Random Forest provides feature importance by checking how much they help organize the data or how much the model fails if that feature is not there.

SHAP-based Explainability

We make use of SHAP (SHapley Additive exPlanations) todetermine feature importance, providing a simple framework helping us in describing predictions across all model types.

- **Theoretical Foundation:** SHAP values are based on Shapley values from cooperative game theory. For a prediction f (x), SHAP assigns each feature i a value ϕi representing its share in the final output of prediction:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

where ϕ0 is the base value (expected model result over the training data) and ϕi is the Shapley value for feature i.
The Shapley value for feature i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$$

where N is the set of all features, S is a subset of features, and f (S) is the model prediction using only features in S.

SHAP Explainers: Different SHAP explainers are en- hanced for different model types:

- **LinearExplainer:** For linear models (LR, SVM), it calculate exact SHAP values optimally using the model coefficients:

$$\phi_i = \beta_i(x_i - \mathbb{E}[x_i])$$

- **TreeExplainer:** For tree-based models (RF), it determines exact SHAP values in polynomial time using a specialized algorithm that makes use of the tree structure.
- **KernelExplainer:** A model-agnostic approach that approximates SHAP values using weighted linear regression. Used for models that don't have specialized explainers.
- **Aggregating Feature Importance:** For a set of n samples, we determine the mean absolute SHAP value for each feature:

$$\bar{\phi}_j = \frac{1}{n}\sum_{i=1}^{n} |\phi_j^{(i)}|$$

Features are ranked by $\bar{\phi}_j$ to determine the most prominent features for fake news detection.

Trust Gap Analysis

We introduce the Trust Gap framework to Measure agreement between different models explaination. The key insight is that models with similar accuracy may depend on different features at times, and understanding this difference is helpful for model selection and ensemble design.

Motivation: Consider two models M1 and M2 that both achieve 95% accuracy on fake news detection. If M1 depends majorly on political keywords while M2 depends on writing style features, which model should we trust? The Trust Gap analysis provides a key approach to answer this question.

Algorithm 3 Trust Gap Analysis

Algorithm 3 Trust Gap Analysis

Require: Models M_1, M_2, SHAP values S_1, S_2, top-k parameter
Ensure: Jaccard similarity J, Rank correlation ρ
{Extract top-k features by mean —SHAP—}
$F_1 \leftarrow$ top_k_features(S_1, k)
$F_2 \leftarrow$ top_k_features(S_2, k)
{Compute Jaccard Similarity}
$J \leftarrow \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$
{Compute Rank Correlation for common features}
$common \leftarrow F_1 \cap F_2$
if $|common| \geq 3$ **then**
  $ranks_1 \leftarrow$ get_ranks$(common, S_1)$
  $ranks_2 \leftarrow$ get_ranks$(common, S_2)$
  $\rho \leftarrow$ spearman_correlation$(ranks_1, ranks_2)$
else
  $\rho \leftarrow$ undefined
end if
return J, ρ

2) Trust Gap Metrics: Jaccard Similarity measures the overlap between top-k features of two models:

$$J(F_1, F_2) = \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$$

A Jaccard similarity of 1.0 means perfect agreement (among identical top features), while 0.0 indicates no overlap. In practice, we find that:

- **J > 0.7:** High agreement - models depends on similar features
- **0.4 < J ≤ 0.7:** Moderate agreement - some shared features
- **J ≤ 0.4:** Low agreement - models use different features entirely

Spearman Rank Correlation measures whether common features are ranked likewise by both models:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where di is the difference in ranks for feature i between the two models, and n is the number of common features.

A positive ρ indicates that features relevant to one model are also relevant to the other. A negative ρ suggests opposite relationship in feature importance.

- **Interpreting the Trust Gap:** The Trust Gap has practical Conclusions:
- **High Trust Gap (low J):** Models do not agree on important features. This may means that one model has learned fake correlations, or that multiple valid detection strategies exist.
- **Low Trust Gap (high J):** Models agree on important features. This increases confidence that the identified features are real indicators of fake news.

G. Cross-Dataset Robustness Analysis

To analyze feature generalizability across domains, we compare SHAP-derived feature importance between datasets.

- **Motivation:** A fake news detection model trained on political news may learn features specific to political realm (e.g., names of politicians, political parties). When applied to health misinformation, these features become useless. Understanding which features are universal versus topic-specific is crucial for developing reliable system.

Algorithm 4 Cross-Dataset Feature Analysis

Require: Top-k features from Dataset A: F_A, Dataset B: F_B
Ensure: Universal features U, Topic-specific features T_A, T_B, Universality Ratio R

$U \leftarrow F_A \cap F_B$ {Features important in both datasets}
$T_A \leftarrow F_A \setminus F_B$ {Features specific to Dataset A}
$T_B \leftarrow F_B \setminus F_A$ {Features specific to Dataset B}
$R \leftarrow \frac{|U|}{k}$ {Universality Ratio}
return U, T_A, T_B, R

Cross-Dataset Analysis Framework:

Universality Ratio: The Universality Ratio measures what part of important features are applicable across different realm:

where k is the number of top features considered. We determine:

- **R > 0.5**: Major part of features are universal - good cross-domain potential
- **0.25 < R ≤ 0.5:** Mixed - some universal, some topic-specific features
- **R ≤ 0.25:** Mostly topic-specific features - poor cross-domain potential

4) Feature Categories: Our analysis divides features into three groups:

- **Universal Features:** Important in both datasets. Theselikely capture basic characteristics of fake news (e.g., sensationalist language, lack of attribution)
- **Topic-Specific Features (Dataset A):** Important only in Dataset A. For political news, these might include politician names or political terms.
- **Topic-Specific Features (Dataset B):** Important only in Dataset B. For general news, these might include entertainment or lifestyle terms.

IV. EXPERIMENTAL SETUP

A. Datasets

We used 2 available datasets which represent different news domains, selected for their size, quality, and diversity.

1) Dataset A - ISOT Fake News Dataset: The ISOT dataset [17] contains political news articles from 2015 to 2018. Real news articles were fetched from Reuters.com, a reputed international news organization. Fake news articles were collected from various fake-news websites flagged by fact-checking organizations including PolitiFact and Wikipedia's list of fake news websites.

- Total articles: 44,898
- Real news: 21,417 (47.7%)
- Fake news: 23,481 (52.3%)
- Domain: Political news (US politics)
- Time period: 2015-2018
- Average article length: 400 words

2) Dataset B - WELFake Dataset: The WELFake dataset [18] is a large dataset combining news from various sources including Kaggle, McIntire, Reuters and BuzzFeed Political. It covers general news topics apart from politics.

- Total articles: 72,134
- Real news: 35,028 (48.6%)
- Fake news: 37,106 (51.4%)
- Domain: General news (mixed topics)
- Sources: Multiple aggregated sources
- Average article length: 350 words

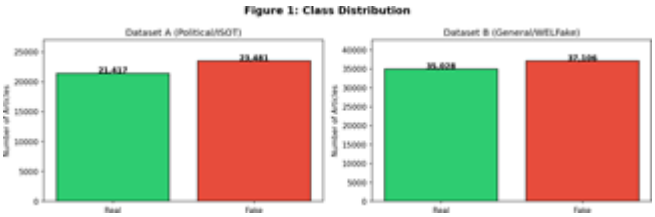

Fig. 1. Class distribution of the two datasets. Both datasets are approximatelybalanced between real and fake news, with slight majority of fake news in each case.

Dataset Comparison: The two datasets differ in several important ways:

- Domain specificity: ISOT focuses on political news, while WELFake covers diverse topics
- Source diversity: ISOT has a single real news source (Reuters), while WELFake aggregates multiple sources
- Size: WELFake is approximately 60% larger than ISOT These differences make the datasets ideal for cross-domain analysis, as features that appear in both are likely to be genuinely indicative of fake news rather than topic-specific artifacts.

B. Implementation Details

All experiments were implemented in Python 3.10 using the following libraries:

- **scikit-learn 1.2:** For machine learning models and TF-IDF vectorization

- **SHAP 0.42**: For computing feature explanations
- **NLTK 3.8:** For text preprocessing and stopword removal
- **pandas 1.5:** For data manipulation
- **matplotlib/seaborn:** For visualization

Hyperparameters:

- **Train-Test Split:** 80%-20% with stratified sampling to maintain class balance
- **TF-IDF Features:** Maximum 5,000 features, minimum document frequency of 2
- **Logistic Regression:** L2 regularization, max iterations = 1000
- **SVM:** Linear kernel, C = 1.0, max iterations = 2000
- **Random Forest:** 100 trees, no maximum depth
- **SHAP Samples:** 200 samples for explanation computation
- **Top-k Features:** k = 20 for Trust Gap and Cross-Dataset analysis
- **Random Seed:** 42 for reproducibility

Computational Resources: Experiments were conducted on [specify your hardware, e.g., "a machine with Intel Core i7 processor, 16GB RAM, running Windows 11"]. Training times ranged from approximately 30 seconds (Logistic Regression) to 5 minutes (Random Forest) per dataset. SHAP computation required approximately 2-10 minutes per model depending on the explainer type.

C. Evaluation Metrics

1) Classification Performance Metrics:

- **Accuracy**: Proportion of correct predictions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** Proportion of predicted fake news that is actually fake
- **Recall:** Proportion of actual fake news that is correctly identified

F1-Score: Harmonic mean of precision and recall

- **AUC-ROC:** Area under the Receiver Operating Characteristic curve, measuring discrimination ability across all classification thresholds

2) Explainability Metrics:

- **Jaccard Similarity:** Feature overlap between models (range: 0-1)
- **Spearman Rank Correlation:** Rank agreement for common features (range: -1 to 1)
- **Universality Ratio:** Cross-dataset feature transferability (range: 0-1)

V. RESULTS AND DISCUSSION

This section presents our experimental results, organized around three research questions:

- **RQ1:** How do different ML models perform on fake news detection?
- **RQ2:** Do different models agree on which features indicate fake news? (Trust Gap)
- **RQ3:** Do fake news patterns generalize across domains? (Cross-Dataset Robustness)

A. **RQ1:** Classification Performance

Table I presents the classification performance of all models on both datasets.

TABLE I
MODEL PERFORMANCE COMPARISON

Dataset	Model	Acc	Prec	Rec	F1	AUC
ISOT	Logistic Reg.	0.990	0.993	0.987	0.990	0.999
	Naive Bayes	0.939	0.941	0.943	0.942	0.985
	SVM	0.995	0.997	0.993	0.995	0.999
	Random Forest	0.997	0.999	0.996	0.998	0.999
WELFake	Logistic Reg.	0.949	0.947	0.955	0.951	0.989
	Naive Bayes	0.844	0.839	0.864	0.851	0.924
	SVM	0.953	0.950	0.959	0.954	0.990
	Random Forest	0.959	0.952	0.970	0.961	0.992

- **Performance Analysis:** All models achieved very good classification performance on both the datasets, with accuracy scores from 84.4% to 99.7%. Key observations include:
- Random Forest achieved the highest performance on both datasets (99.7% accuracy on ISOT, 95.9% on WELFake), which explains the impact of ensemble meth- ods for fake news detection.
- SVM performed neck to neck with Random Forest, achieving 99.5% on ISOT and 95.3% on WELFake,

while offering faster training time and more explainable decision boundaries.

- Logistic Regression achieved 99.0% on ISOT and 94.9% on WELFake, providing a strong baseline with fully explainable coefficients.
- Naive Bayes showed the weakest performance (93.9% on ISOT, 84.4% on WELFake), probably because of the fact that natural language data violated its strong independence assumptions.
- Performance was consistently higher on ISOT (political news) compared to WELFake (general news), suggesting that political fake news may have more distinctive language patterns.

2) Key Observations:

1) All models achieve high accuracy on both the datasets (84-99%), which explains that classical ML approaches are highly competitive for fake news detection even compared to deep learning methods.
2) Random Forest achieves the highest F1-score (0.998 on ISOT, 0.961 on WELFake), which suggests that it has the best balance between precision and recall through its ensemble of decision trees.
3) Performance on ISOT is consistently 3-10% higher than WELFake across all the models, most probably due to ISOT's more homogeneous source (Reuters for real news) creating clearer stylistic distinctions.
4) The precision-recall trade-off is well-balanced across models, with no significant bias toward either metric, indicating robust classification without systematic over- or under-prediction of fake news.

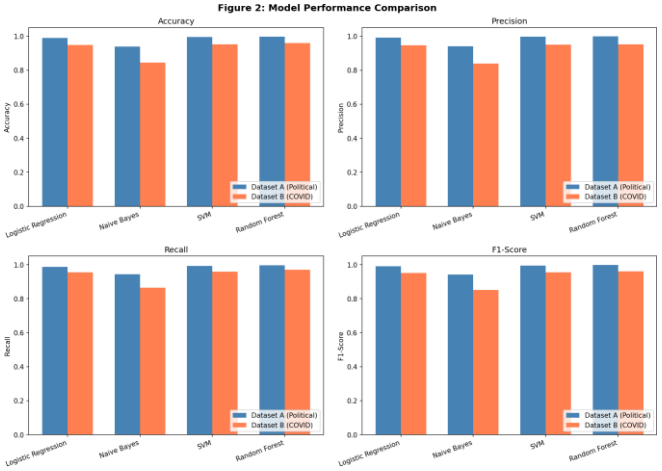

Fig. 2. Performance comparison across all models and datasets. The groupedbar chart shows Accuracy, Precision, Recall, and F1-Score for each model onboth ISOT (political) and WELFake (general) datasets

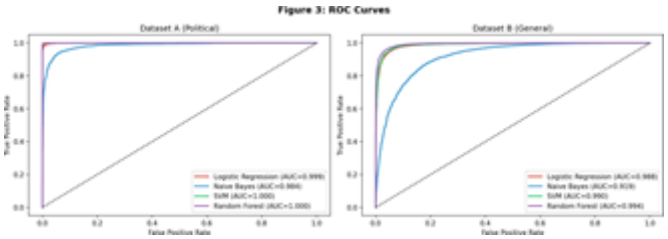

Fig. 3. ROC curves for all models on both datasets. The curves illustrate the trade-off between true positive rate and false positive rate at various classification thresholds. AUC values are shown in the legend.

B. RQ2: SHAP Feature Importance Analysis

Before analyzing the Trust Gap, we first examine the features identified as important by each model.

- **Feature Analysis:** Our SHAP analysis revealed different patterns of how different models identify fake news indicators.

Top Features by Model (ISOT - Political News):

- Logistic Regression: "said", "reuters" "video", "trumps", "us"
- SVM: "reuters", "said", "video", "trumps", "hillary"
- Random Forest: "via", "gun", "daily", "facebook", "americans"

Common Fake News Indicators:

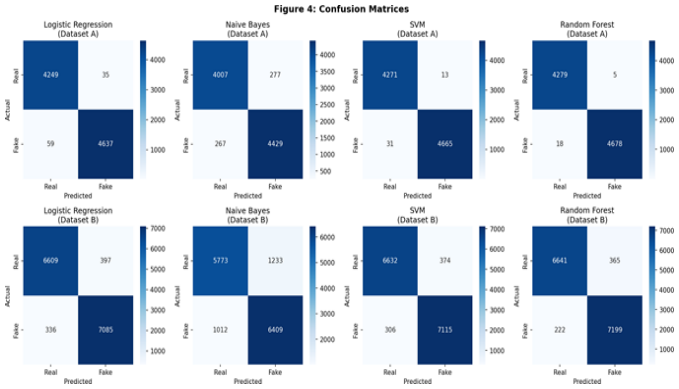

Fig. 4. Confusion matrices for all models across both datasets. Each matrix shows the distribution of true positives, true negatives, false positives, and false negatives.

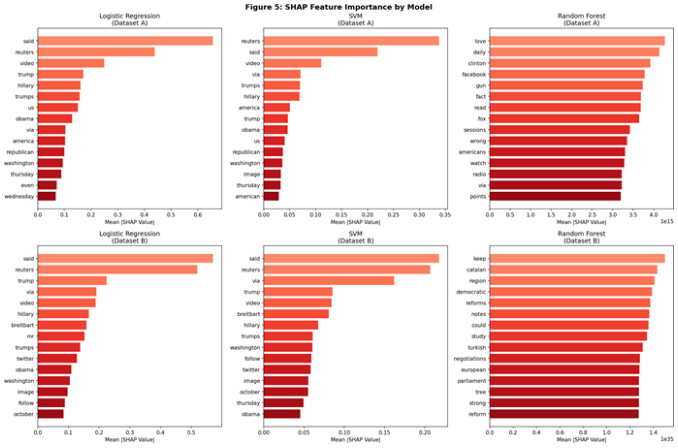

Fig. 5. SHAP feature importance for all models across both datasets. Top 15 features are shown for each model-dataset combination, ranked by mean absolute SHAP value.

- Words like “video”, “via”, and “facebook” suggest fake news often comes from social media content or viral material
- Political terms (“trumps”, “hillary”) appear very much, showing the political nature of the ISOT dataset
- Informal language patterns distinguish fake from real news

Common Real News Indicators:

- “reuters” strongly indicates real news, as it represents professional journalism attribution
- “said” appears frequently in real news, reflecting proper quotation and sourcing practices
- Formal reporting language (“washington”, “house”, “senate”) characterizes legitimate political journalism

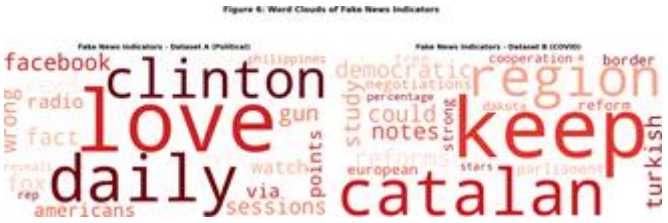

Fig. 6. Word clouds visualizing fake news indicator words for each dataset based on Random Forest SHAP values. Larger words have higher importance scores.

C. RQ2: Trust Gap Analysis

Table II presents the Jaccard similarity between model pairs, quantifying the Trust Gap.

TABLE II

TRUST GAP ANALYSIS: JACCARD SIMILARITY BETWEEN MODELS (TOP-20 FEATURES)

Model Pair	ISOT	WELFake
LR vs SVM	0.739	1.000
LR vs RF	0.053	0.000
SVM vs RF	0.053	0.000
Average	0.282	0.333

TABLE III

Trust Gap Analysis: Spearman Rank Correlation For Common Features

Model Pair	ISOT	WELFake
LR vs SVM	0.907	0.836
LR vs RF	N/A	N/A
SVM vs RF	N/A	N/A

- **Trust Gap Findings:** Our Trust Gap analysis reveals striking differences in how models approach fake news de- tection.

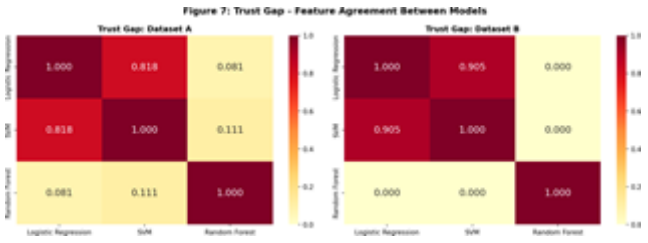

Fig. 7. Trust Gap heatmaps showing Jaccard similarity between model explanations for both datasets. Higher values (darker colors) indicate greater agreement on which features indicate fake news.

Key Finding 1: Linear Models Show High Agreement
Logistic Regression and SVM exhibit remarkably high agreement, with Jaccard similarity of 0.739 on ISOT and perfect agreement (1.000) on WELFake. The Spearman rank correlation of 0.907 (ISOT) and 0.836 (WELFake) confirms that not only do they share features, but they rank them similarly. This is expected since both are linear models optimizing similar objectives—they identify the same linear separating hyperplane in feature space.

Common features between LR and SVM include: “said”, “reuters”, “washington”, “video”, “even” (ISOT) and “said”, “image”, “video”, “monday”, “obama” (WELFake).

Key Finding 2: Random Forest Uses Fundamentally Different Features

The most striking finding is the near-zero agreement between Random Forest and linear models (Jaccard = 0.053 on ISOT, 0.000 on WELFake). This represents a significant "Trust Gap"—despite all models achieving similar accuracy, Random Forest relies on entirely different textual patterns.

On ISOT, RF shares only "via" and "trump" with linear models. On WELFake, there is zero overlap in top-20 features. This suggests RF captures non-linear interactions and different linguistic patterns that linear models cannot detect.

Key Finding 3: Trust Gap is Dataset-Dependent

The Trust Gap between linear models is lower on WELFake (perfect agreement) than ISOT (0.739), suggesting that general news fake detection may have more consistent patterns across model types, while political news allows for more diverse detection strategies.

D. RQ3: Cross-Dataset Robustness

Table IV shows the universality ratio for each model, indicating what proportion of top-20 features transfer acrossdatasets.

TABLE IV
CROSS-DATASET FEATURE ANALYSIS

Model	Universal	ISOT-only	WELFake-only	Ratio
Logistic Reg.	10	10	10	50%
SVM	11	9	9	55%
Random Forest	1	19	19	5%
Average	7.3	12.7	12.7	37%

Universal Features Analysis: Our cross-dataset analysis identified features that generalize across political and general news domains.

Universal Fake News Indicators (present in both datasets):

Fig. 8. Cross-dataset feature comparison showing top features for each model across both datasets. Features appearing in both datasets are highlighted.

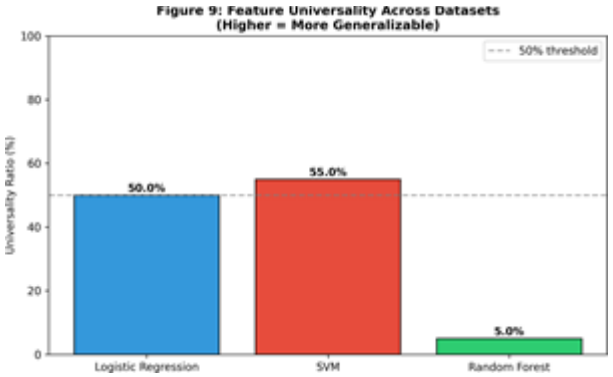

Fig. 9. Universality ratio across models. Higher values indicate that a larger proportion of important features generalize across domains.

- "said": Appears in both datasets as a key indicator. Interestingly, proper attribution ("X said") is more common in real news, while fake news often lacks clear sourcing.
- "reuters": A strong real news indicator, as Reuters attribution signals professional journalism.
- "washington": Geographic references to news centers appear in legitimate reporting.
- "video" and "image": References to multimedia content appear across domains, often in fake news that relies on visual clickbait.
- "hillary": Political figure names transfer across datasets, as political misinformation spans over multiple news categories.

These universal features capture fundamental characteristics of fake news that transcend specific topics:

- Lack of proper attribution or sourcing (absence of "said", "reuters")
- Reliance on multimedia references rather than substantive reporting
- Informal writing style compared to professional journalism

2) Topic-Specific Features Analysis: ISOT-Specific Features (Political News):

- "house", "senate", "republican": US political institution and party references

- "wednesday", "us": Temporal and geographic markers specific to US political reporting
- "even": Intensifier common in political commentary

These features reflect the political focus of ISOT and wouldn't transfer to non-political domains.

WELFake-Specific Features (General News):

- "breitbart": Source-specific indicator (Breitbart articles in the dataset)
- "friday", "november", "monday": Temporal markers more prominent in general news
- "mr": Formal title usage patterns differ between datasets
- "twitter": Social media references more prominent in general news fake content

These features are artifacts of the specific sources and time periods in WELFake.

3) Implications for Cross-Domain Deployment: The average universality ratio of 37% (50-55% for linear models, only 5% for Random Forest) reveals important insights for deployment:

1. **Model Retraining:** Linear models (LR, SVM) with 50-55% universality can partially transfer to new domains but will benefit from fine-tuning. Random Forest with only 5% universality requires complete retraining for new domains.
2. **Feature Engineering:** Practitioners should focus on universal features (attribution patterns, sourcing language) when building generalizable systems, while accepting that some domain-specific features will need to be learned per-domain.
3. **Transfer Learning:** The moderate universality ratio (50-55%) for linear models suggests that transfer learning approaches would be partially effective—pre-training on one domain provides a useful starting point but domain adaptation remains necessary.
4. **Model Selection Trade-off:** Random Forest achieves highest accuracy but lowest generalizability. For cross-domain applications, linear models offer a better accuracy-generalizability trade-off.

Discussion

1. **Implications for Model Selection:** Our Trust Gap analysis reveals that different models, despite achieving similar accuracy, may rely on fundamentally different features. This has several implications:
2. **Ensemble Design:** Models with low Trust Gap (high agreement) may provide redundant information in ensembles. Combining models with moderate Trust Gap could capture complementary patterns.
3. **Explanation Consistency**: For applications requiring explanations (e.g., content moderation), practitioners should prefer models that agree with other approaches, as this increases confidence in the explanations.
4. **Robustness Assessment**: High Trust Gap may indicate that some models have learned spurious correlations. Cross-validating explanations across models can help identify such issues.
5. **Relationship Between Trust Gap and Cross-Domain Ro- bustness:** An interesting finding is the inverse relationship between Trust Gap and cross-domain robustness. Models with high inter-model agreement (LR and SVM, Jaccard 0.739- 1.000) also exhibit higher universality ratios (50-55%), while Random Forest shows both high Trust Gap (near-zero agree- ment with linear models) and poor cross-domain transfer (5% universality).

This suggests that consensus features—those identified by multiple model architectures—are more likely to represent genuine, generalizable fake news indicators rather than dataset- specific artifacts. The features that linear models agree upon ("said", "reuters", "washington") capture fundamental journal- istic patterns, while Random Forest's unique features ("via", "gun", "daily", "facebook") may represent dataset-specific correlations that do not transfer.

Practical Recommendations: Based on our findings, we offer the following recommendations for practitioners:

- **For High-Stakes Applications:** Use models with low Trust Gap and high universality ratio to ensure consistent and generalizable detection.
- **For Domain-Specific Deployment:** If deploying within a specific domain (e.g., only political news), models can leverage topic-specific features for higher accuracy.
- **For Cross-Domain Deployment:** Focus on universal features and consider retraining or fine-tuning when moving to new domains.
- **For Explainability**: Provide explanations from multiple models to users, highlighting areas of agreement and disagreement.

Limitations: Our study has several limitations that should be considered:

1) **Model Scope:** Our analysis is limited to classical ML models. Deep learning models (BERT, GPT) may exhibit different trust gap patterns and should be checked in future.

2) **SHAP Approximations:** While SHAP provides theoretically grounded explanations, the values are approximate and may not capture all aspects of model behavior, specially for complex non-linear models.

3) **Dataset Limitations:** Although we use two diverse datasets, they may not represent all fake news do- mains (e.g., health misinformation, financial fraud). The datasets are also mostly in English.

4) **Temporal Aspects:** Fake news patterns evolve over time. Our analysis uses static datasets and does not capture chronological patterns.

5) Feature Representation: TF-IDF captures lexical patterns but may miss semantic refinements. Word embed- dings or contextual representations might show different patterns.

Threats to Validity:

- **Internal Validity:** Random seed selection and train- test splits could affect results. We mitigate this through stratified sampling and consistent random seeds.
- **External Validity**: Results may not generalize to other languages, time periods, or fake news types not represented in our datasets.
- **Construct Validity:** Jaccard similarity and Spearman correlation are reasonable but not the only ways to measure model agreement. Alternative metrics might yield different insights.

VI. CONCLUSION

This paper present's a extensive model for comparative and explainable fake news detection, addressing the crucial need for transparency and Stability in automated misinformation detection systems.

A. Summary of Contributions

1. Trust Gap Analysis Framework

We introduced a innovative approach to assess agreement among different ML model explanations using SHAP values, Jaccard similarity, and Spearman rank correlation. Our Examination of four classical ML models on two datasets revealed that:

- Linear models (LR, SVM) show high agreement on the features (Jaccard 0.739-1.000), while Random Forest uses different features altogether(Jaccard 0.000-0.053 with linear models)
- because model struggle when data source don't align , depending on shared characteristic's results in more valid results.
- Linear models (LR, SVM) show higher agreement as compared to ensemble methods (RF), making them more suitable for applications that need uniform explanations

The Trust Gap framework give the practitioners a principled approach to evaluate model consistency and make informed decisions about which models to deploy in production systems to yield optimal results.

1. Cross-Dataset Robustness Analysis

We systematically analyzed feature generalizability across realms, identifying:

- Universal features: said, reuters, washington, video , hillary that generalize across political and general news domains
- Topic-specific features: Domain-dependent indicators like house, senate (political) and breitbart, twitter (general) that require retraining when deploying to new domains
- An average universality ratio of 37% (50-55% for linear models, 5% for Random Forest), suggesting moderate cross-domain potential for linear models but poor generalizability for tree-based ensembles

2. Practical Insights

Our findings provide practical guidance for developers:

- For explainable systems, prefer models with low Trust Gap to ensure uniform explanations
- For cross-domain deployment, focus on universal features and plan for domain adaptation
- Consider ensemble methods that combine models with complementary feature sets

B. Broader Impact

Fake news detection systems have notable social consequences. Our work contributes to more honest and reliable systems by:

- Enabling users to understand why content is classified as potentially fake

- Helping developers identify when models may be learning fake patterns in the dataset
- Providing insights into the basic characteristics of fake news that are universal across different topics

However, we know that automated fake news detection is not a complete solution. Human judgment, media literacy, and institutional fact-checking remain as an important components of tackling the spread of fabricated content.

C. Future Work

Several promising factors emerge from this research for future work:

1. **Deep Learning Extension:** Apply the Trust Gap framework to Transformer-based models (BERT, RoBERTa, GPT) to understand how attention mechanisms identify fake news patterns.
2. **Multimodal Analysis:** Extend the framework to incorporate images, videos, and social context features, analyzing which Media types contribute most to detection.
3. **Temporal Dynamics:** Examine how fake news patterns and model explanations evolve with time, particularly during major events (elections, pandemics).
4. **Adversarial Robustness:** Use Trust Gap insights to develop more reliable models that are immune against Conflicting manipulation.
5. **User Studies:** Conduct human-subject experiments to find out whether this Trust Gap-informed explanations help in improving user understanding and trust in fake news detection systems.
6. **Multilingual Extension:** Apply the framework to other datasets than english to understand common patterns of misinformation in multiple languages.
7. **Real-time Systems:** Develop optimal implementations suitable for real-time fake news detection in social media platforms.

Acknowledgment

We are deeply indebted to Rajat Thakkar from Chitkara University for his mentorship. We are grateful for the time he invested in reviewing our progress and for the professional wisdom he shared, which significantly enhanced the quality of this project.

Ethical Considerations

This research was conducted using publicly available datasets. We acknowledge that fake news detection systems can be misused for censorship. We advocate for transparent deployment with human oversight and appeal mechanisms.

REFERENCES

1. S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.
2. J. Zarocostas, “How to fight an infodemic,” The Lancet, vol. 395, no. 10225, p. 676, 2020.
3. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
4. X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities,” ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020.
5. C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on Twitter,” in Proc. WWW, 2011, pp. 675–684.
6. H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using n-gram analysis and machine learning techniques,” in Proc. ISDDC, 2017, pp. 127–138.
7. W. Y. Wang, “Liar, liar pants on fire: A new benchmark dataset for fake news detection,” in Proc. ACL, 2017, pp. 422–426.
8. Y. Wang et al., “EANN: Event adversarial neural networks for multi- modal fake news detection,” in Proc. KDD, 2018, pp. 849–857.
9. J. Ma et al., “Detecting rumors from microblogs with recurrent neural networks,” in Proc. IJCAI, 2016, pp. 3818–3824.
10. R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach,” Multimedia Tools and Applications, vol. 80, pp. 11765–11788, 2021.
11. A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Con- cepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, 2020.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in Proc. KDD, 2016, pp. 1135–1144.
13. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. NeurIPS, 2017, pp. 4765–4774.

14. J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, “Ex- plainable machine learning for fake news detection,” in Proc. WebSci, 2019, pp. 17–26.
15. K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “dEFEND: Explainable fake news detection,” in Proc. KDD, 2019, pp. 395–405.
16. R. M. Silva, R. L. Santos, T. A. Almeida, and T. A. Pardo, “Towards automatically filtering fake news in Portuguese,” Expert Systems with Applications, vol. 146, p. 113199, 2020.
17. H. Ahmed, I. Traore, and S. Saad, “Detecting opinion spams and fake news using text classification,” Security and Privacy, vol. 1, no. 1, p. e9, 2018.
18. P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, “WELFake: Word embedding over linguistic features for fake news detection,” IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881–893, 2021.
19. B. D. Horne and S. Adali, “This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news,” in Proc. ICWSM, 2017, pp. 759–766.
20. V. Pe´rez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” in Proc. COLING, 2018, pp. 3391–3401.
21. H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact- checking,” in Proc. EMNLP, 2017, pp. 2931–2937.
22. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A stylometric inquiry into hyperpartisan and fake news,” in Proc. ACL, 2018, pp. 231–240.
23. V. L. Rubin, Y. Chen, and N. J. Conroy, “Deception detection for news: Three types of fakes,” in Proc. ASIS&T, 2015, pp. 1–4.
24. N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” in Proc. ASIS&T, 2015, pp. 1–4.
25. X. Zhang and A. A. Ghorbani, “An overview of online fake news: Characterization, detection, and discussion,” Information Processing & Management, vol. 57, no. 2, p. 102025, 2020.